Received: 25 March 2020 | Revised: 7 July 2020 | Accepted: 14 July 2020

DOI: 10.1002/widm.1385

OVERVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

# Combating disinformation in a social media age

**Kai Shu**[1] | **Amrita Bhattacharjee**[2] | **Faisal Alatawi**[2] | **Tahora H. Nazer**[3] | **Kaize Ding**[2] | **Mansooreh Karami**[2] | **Huan Liu**[2]

[1]Department of Computer Science, Illinois Institute of Technology, Chicago, Illinois

[2]Computer Science and Engineering, Arizona State University, Tempe, Arizona

[3]Spotify, Boston, Massachusetts

**Correspondence**
Kai Shu, Department of Computer Science, Illinois Institute of Technology, Chicago, IL.
Email: kshu@iit.edu

**Funding information**
National Science Foundation, Grant/Award Numbers: 1614576, 1909555

**Abstract**

The creation, dissemination, and consumption of disinformation and fabricated content on social media is a growing concern, especially with the ease of access to such sources, and the lack of awareness of the existence of such false information. In this article, we present an overview of the techniques explored to date for the combating of disinformation with various forms. We introduce different forms of disinformation, discuss factors related to the spread of disinformation, elaborate on the inherent challenges in detecting disinformation, and show some approaches to mitigating disinformation via education, research, and collaboration. Looking ahead, we present some promising future research directions on disinformation.

This article is categorized under:
Algorithmic Development > Multimedia
Commercial, Legal, and Ethical Issues > Social Considerations
Application Areas > Education and Learning

**KEYWORDS**

disinformation, fake news, misinformation

## 1 | INTRODUCTION

The proliferation and prevalence of social media in almost every facet of human lives have made the consumption of news and information extremely convenient to the users of such technology. The ease with which information, alerts, and warnings can be broadcasted to millions of people in a very short amount of time, has made social media a brilliant platform for information diffusion, especially for time-sensitive situations, for example, during natural disasters and crisis events. Given that a considerable fraction of individuals across the world use social media as a source of news (Nami Sumida & Mitchell, 2019; Shearer & Matsa, 2018), and thus letting such news affect their opinions and actions, directly and indirectly, checking the veracity of such news becomes an important task.

Over the last few years, the existence of disinformation online and the malicious agents acting as sources of such disinformation has been recognized and acknowledged. Research in the domain of disinformation detection and related fields has gained momentum, and different aspects of the problem are being approached by researchers from different perspectives.

In this article, we provide a comprehensive overview of the past and ongoing research in disinformation. We begin by defining a few relevant terms used extensively in the literature. In the following sections, we look at the history of disinformation and its characteristics in the social media age. Then we proceed to explain the challenges in the detection of disinformation, the different forms of disinformation that are prevalent on social media, and detection and

  

mitigation approaches and techniques. We further talk about the factors behind the rapid spread of disinformation, steps taken to educate people about online disinformation, and conclude the review with some possible areas of future work.

## 1.1 | Definitions

Upon a thorough review of existing literature in the context of the detection of deliberately fabricated content, we frequently come across the following keywords and terminologies—*Misinformation*, *Disinformation*, *Fake News*, *Hoax*, *Rumor*, and *Conspiracy theory*. In this section, we proceed to define these terms in the way these have been used by most researchers.

*Misinformation* (Wu, Morstatter, Carley, & Liu, 2019) has been described in the literature as "false, mistaken or misleading" information (Fetzer, 2004), often considered as an "honest mistake."

On the other hand, *disinformation* is false information, spread deliberately with the intention to mislead and/or deceive (Hernon, 1995). *Fake news* has been defined as "news articles that are intentionally and verifiably false, and could mislead readers" (Allcott & Gentzkow, 2017), and most researchers follow this definition. So fake news is an example of disinformation. In this article, we use the terms "disinformation" and "fake news" interchangeably.

Many researchers also use the term "hoax" (originating from *hocus*, which means to trick or deceive), to refer to deliberate false information (Kumar, West, & Leskovec, 2016; Pratiwi, Asmara, & Rahutomo, 2017; Santoso, Yohansen, Nealson, Warnars, & Hashimoto, 2017; Vedova et al., 2018). Hoaxes are messages created with the intention of being spread to a large number of people, to "persuade or manipulate other people to do or prevent pre-established actions, mostly by using a threat or deception" (Hernandez, Hernandez, Sierra, & Ribagorda, 2002; Vuković, Pripužić, & Belani, 2009).

There has also been some work in the modeling of the spread of "rumors." According to (Rosnow, 1991), "rumors" are "public communications that are infused with private hypotheses about how the world works."

Although we do not cover it in much detail, "conspiracy" and "conspiracy theory" are also terminologies that researchers have used in related works. According to van der Tempel and Alcock (2015), conspiracy theories are beliefs that are largely disregarded by society. A conspiracy theory "involves not only the denial of official, commonly-held explanations about the causes of an event, but also the attribution of the event to a plan devised by a group of agents with hidden, unlawful, and malevolent intent". For details on how conspiracy theories spread and government responses so far, we direct readers to Sunstein and Vermeule (2009).

## 1.2 | A brief history of disinformation

The existence of disinformation and fake news is not new (Taylor, 2016), and the term has been widely in use since 1950s (Manning, Manning, & Romerstein, 2004). For several decades individuals, groups and governments have tried to tarnish public opinion by exposing them to falsified, forged information as a way to sway people's political alignment. Deception and disinformation have been used by political forces for decades to have an edge over opponents/enemies—one of the famous documented instances of deception that succeeded in its mission was Operation Bodyguard, planned and executed by the Allied forces during World War II. The years following World War II, including the Cold War era, saw frequent use of forged, falsified information, clandestine radio stations, and newspaper articles to mislead the public and also governments. Several instances of such deceit and forgery carried out by governments have been documented in reports published by the CIA and the United States Bureau of Public Affairs (Department of State, 1981). Another infamous historical disinformation campaign was the Operation INFEKTION, active in the 1980s, which misled the public to believe that HIV/AIDS was developed by the US as a biological weapon (Boghardt, 2009).

Instances of political incidents where disinformation was prevalent, thus affecting public sentiment and opinions, include the assassination of president John F. Kennedy. Fetzer (2004) explains his work on this issue and the kinds of disinformation he encountered in documents released on the president's death. He further talks about studies where a movie of the assassination has been questioned, and claims have been made that the movie has been doctored—several frames have been edited out and the sequence of frames has been changed.

## 1.3 | Disinformation in the age of social media

While the advent of the Internet came as a boon to society, and its gradual adoption has resulted in a more connected world, the reachability of the internet also means it could be misused successfully. The Internet provides a platform for fast dissemination of information and has often been referred to as the "information highway" in the literature (De Maeyer, 1997). The positive impact of this is profound—with individuals becoming more aware of local and global news, their rights, raising awareness regarding global concerns including climate change and plastic pollution—which ultimately resulted in movements and aimed at taking action. However, all of these are based on the unifying assumption that the information available to people is real and not designed to mislead.

The connectivity of the Internet has been misused as well, to spread propaganda, to fulfill some individual or group agenda. There have been countless incidents of human lives being at stake due to false sensationalized information being spread via social media. Hoax messages and rumors on messaging platforms like WhatsApp and Facebook spread like wildfire, and with an increase in smartphone usage, this makes the gullible public treat these false claims as genuine information and accordingly form opinions and take actions. Recently, there have been several reported instances of vigilante justice where innocent people were lynched to death by an infuriated mob after rumors about the victims committing some kind of crime went viral. Social media platforms like WhatsApp and Facebook have been targeted by miscreants to spread rumors and hoaxes (Parth & Bengali, 2018). Some brutal incidents include one where a 21-year-old innocent boy in a Mexican village was beaten and burned to death by an enraged mob (Patrick & McDonnell, 2018). Similar cases of lynchings and mob violence based solely on false claims and hoaxes propagated via WhatsApp have been a matter of concern in India, which happens to be the largest market for the company, having about 200 million active monthly users. Since this has become a growing concern, many studies have been conducted with a focus on how to detect and thwart the spread of disinformation and hoaxes (Arun, 2019), particularly on social media (Shu & Liu, 2019; Shu, Sliva, Wang, Tang, & Liu, 2017; Yang et al., 2019), but arriving at a unified solution has been challenging.

## 1.4 | The challenges of detecting disinformation

We can categorize the challenges of detecting disinformation into two categories (a) content-related and (b) user-related challenges. The content of disinformation, in many cases, is highly sensationalized and is written using extreme sentiments, usually to affect the reader, which makes them interact with the post more (Shu et al., 2017). Thus, such posts containing fabricated content often become "viral" and "trending" on social media (Vosoughi, Roy, & Aral, 2018). In addition to that, the low cost of creating disinformation sources and the ease of using software-controlled social media bots to help spread disinformation (Shao, Ciampaglia, Varol, Flammini, & Menczer, 2017). From the user perspective, social media users are susceptible to disinformation, and they often lack awareness of disinformation (Sharma et al., 2019).

Disinformation such as fake news tends to be much more novel than the truth across all novelty metrics (Vosoughi et al., 2018). People like to share novel news because new information is more valuable to people from an information-theoretic perspective. Furthermore, novel news is valuable from a social perspective where people like to project the image of a person who is "in the know," giving them a unique social status (Vosoughi et al., 2018). People tend to trust their social contacts, and "information" that appear to be trendy. Therefore, spreading disinformation needs convincing content that looks like real articles and a method to make this news go viral (Shao et al., 2017).

The goal of the early detection of disinformation is to prevent its further propagation on social media by giving early alerts of disinformation during the dissemination process (Shu & Liu, 2019). Early detection of disinformation is extremely important to minimize the number of people influenced and hence minimize the damage. This becomes a challenge for automatic detection of disinformation as such detectors do not have access to user reactions, responses, or knowledge bases for fact-checking early on, which could have been helpful for detection. However, exposing people to disinformation might help to increase the true positive rate of the disinformation detectors (Kim, Tabibian, Oh, Schölkopf, & Gomez-Rodriguez, 2018). Furthermore, models such as epidemiological models of disease spread (Bettencourt, Cintrón-Arias, Kaiser, & Castillo-Chávez, 2006; Jin, Dougherty, Saraf, Cao, & Ramakrishnan, 2013) and diffusion models like the independent cascade model (Goldenberg, Libai, & Muller, 2001a; Goldenberg, Libai, & Muller, 2001b) and the linear threshold model (Granovetter, 1978; Zafarani, Abbasi, & Liu, 2014) representing the propagation and spread of disinformation have also been explored to understand and potentially mitigate the spread of

disinformation (Nguyen, Yan, Thai, & Eidenbenz, 2012). However, these models can only be adequately studied after a significant amount of disinformation spread has already occurred, and therefore, this approach of detecting fake news/disinformation is not the most viable. Another challenge behind fake news detection is that some keywords and ways of expression are specific to a certain kind of event or topic. So when a fake news classifier is trained on fake versus real articles based on a particular event or topic, the classifier ends up learning the event-specific features and thus may not perform well when applied to classify fake versus real articles corresponding to another kind of event. So, in general, fake news classifiers need to be generalized to be event-independent.

An important aspect of social media platforms that need to be considered is the existence of filter bubbles or echo chambers created as a result of recommender systems on these platforms. Given the abundance of different kinds of content available online, social media platforms aim to use algorithms that would let users view and interact with content that is most relevant to them. This usually means that users would be exposed to posts, articles, and viewpoints that align with their own beliefs, thus reinforcing them in the process, while also remaining unaware of opposing narratives and beliefs (Pariser, 2011). These "echo-chambers" or "filter bubbles" make the task of disinformation detection and mitigation especially difficult. Being exposed to the same viewpoint repeatedly would only reinforce the pre-existing beliefs of readers, and they would resist changing their opinion even if the narrative is later proven to be false by some fact-checking organization.

## 2 | DISINFORMATION IN DIFFERENT FORMS

Disinformation can exist in different forms—false text, images, videos, and others. In addition, there are different ways by which false content can be generated—both by humans and machines or a combination of the two. This makes the detection process more complex because most detection mechanisms assume some model of generation. In the following sections, we describe the most prevalent forms of false content on social media, and some detection techniques that have been explored by researchers.

### 2.1 | GAN generated fake images

Creating fabricated images and editing images into morphed representations can entirely change the semantic meaning of the image, and such doctored images are often propagated on social media. These images can be created by human agents or machine-generated for example, by Generative Adversarial Networks (GANs; Goodfellow et al., 2014). If such computer-generated images are realistic enough, it becomes very difficult for a human viewer to distinguish between real and such artificially generated images. It has been established that detecting GAN generated fake images is more challenging than images manually modified by human agents (Li, Chang, & Lyu, 2018; Tariq, Lee, Kim, Shin, & Woo, 2018). This is because GANs generate a single image as a whole, so conventional techniques of detecting whether an image has been hand-edited or not, like analyzing metadata of the image, analyzing modifications on the frequency domain, finding disparities in JPEG quality of different regions of the same image, and so on. fail when it comes to images generated by GANs. There is considerable ongoing work on the detection of GAN generated images, and we summarize a few of the promising approaches.

Marra, Gragnaniello, Cozzolino, and Verdoliva (2018) have discussed and compared the performance of several fake image detectors on a dataset of images comprising of both real and computer-generated images. The "fake" images were generated using a CycleGAN, described in Zhu, Park, Isola, and Efros (2017), where an image from one domain was translated into an image from a target domain, and a similar inverse translation was done on the transformed image to convert it back to the original image. Because images uploaded on most social networking sites undergo some form of compression, Marra et al. (2018) conducted experiments comparing the fake image detectors on the compressed version of images as well and compared their performance. It was seen that, in the case of compressed images, even with an overall decline in the performance of all the detectors studied, XceptionNet (Chollet, 2017) gave quite accurate results.

While this work provides valuable insight into the efficacy of several image detectors on artificially generated images and images that have undergone compression, it has assumed the use of one particular GAN model for image generation. However, in real-world fake image detection, it is almost impossible to get access to the model used by the attacker, or even guess what the model could have been. Without this prior knowledge, it becomes very difficult to train classifiers to differentiate between fake and real images that would perform well. Zhang, Karaman, and Chang (2019)

have extended concepts from Odena, Dumoulin, and Olah (2016), which says that GAN generated images have "checkerboard artifacts" as a result of the upsampling(deconvolution) layers, and the presence of these artifacts can guide fake image detectors to distinguish between real and fake images. Utilizing these artifacts, the authors in Zhang et al. (2019) have conducted experiments to classify real versus fake GAN generated images, achieving varying degrees of success for images from different domains.

Other related approaches attempted to address this problem of detection of GAN generated images include the use of co-occurrence matrices by Nataraj et al. (2019). Co-occurrence matrices were computed on the image pixels for each of the three color channels with the assumption that co-occurrence matrices for GAN generated fake images would differ significantly from those of original images, and this would enable a fake image detector to learn features from the matrices and hence differentiate between real and fake images. Tariq et al. (2018) have used an ensemble of shallow CNN based classifiers to detect "real" versus "fake" faces generated by both GANs and by humans using professional image processing software.

## 2.2 | Fake videos and deepfakes

Recent advancements in computer vision have enabled generative adversarial networks to learn to replace faces in videos with that of someone else, and ultimately create realistic videos that can easily fool gullible viewers into thinking it is genuine. This technology, also referred to as Deepfake, has been mostly used for malicious purposes—from using celebrities' faces in pornographic videos (Lee, 2018), to creating fake videos of politicians, which has the potential to shift public sentiment and affect election procedures (George, 2019). The advances in the deepfake video generation are being complemented by equally active research in the detection of such artificially generated videos. Here, we summarize a few of the recent notable works.

Deepfake videos generated using GANs often have certain imperfections if not post-processed correctly. For videos with human subjects involved in some kind of activity, the face swap by the GAN is done on a per-frame basis. So the dependency between consecutive frames is not utilized, and hence, can result in discrepancies between frames, for example, in the hue and/or illumination of the face region. The idea of across-frame inconsistencies has been used by Güera and Delp (2018). The authors used a "convolutional LSTM" consisting of a CNN to extract framewise features, which are then used as sequential input to an LSTM, and outputs from the LSTM are used to detect whether the video is a deepfake video or a non-manipulated one.

Li et al. (2018) have attempted to detect fake videos by detecting eye blinking, with the assumption that in an artificially generated video, eye blinking may not be present or may be somewhat abnormal. This is because the models that generate fake videos by replacing someone's face in an already existing video, usually do not have access to images where the face has eyes closed. Another approach based on detecting discrepancies between real and artificially generated videos involves the analysis of 3D head pose in face videos (Yang, Li, & Lyu, 2019). "Facial landmarks" were detected, which are points that map to features on the face—such as eyebrows, nose, mouth (which constitute the central part of the face), and the boundary or edge of the face. Deepfake conversion of a video usually replaces the central part of a face with that of a different face. During this conversion, there might arise some differences in alignment, which is not natural, and these result in inconsistent head poses. These fine misalignments can be difficult for a human to discern but can be learned by a model to distinguish between real and deepfake videos.

## 2.3 | Multimodal content

Most fake news articles and posts that are propagated online are composed of multiple types of data combined together—for example, a fabricated image along with a text related to the image. This is usually done to make the story/article more believable for the user viewing it. Moreover, there might also be comments on the post where some users may have pointed out that the article seems fabricated and fake, thus in a way guiding other users to verify the credibility of the source. This *multimodal* nature of fake news articles is often quite challenging to handle when it comes to detection using deep learning-based methods, and neural network architectures need to be carefully designed to make use of all the data available and the relationships between them.

Given the challenge of developing an event-independent disinformation detection model, Wang et al. (2018) designed a neural network architecture that has an event discriminator that tries to classify each post/article into the event that it is based on. The idea is to maximize the discriminator loss so that the model learns event-invariant features. The authors have addressed the multimodality of data on social media and have used both text and associated images, extracted event-invariant features from these, and used it to train the fake news detector. Their experiments showed that multimodal detectors performed better than single-modality ones, and learning event-invariant features greatly improved performance. Furthermore, their model performed significantly better than several state-of-the-art for multimodal fake news detection.

A similar but simpler approach using both image and text in news articles was explored by Yang et al. (2018). The authors worked on a fake news dataset containing news articles collected from online sources, and used CNNs to extract features for both text and corresponding image input. Their experimental results also strengthen the idea that incorporating both image and text, and utilizing the multimodal nature of news, provides more information to fake news detectors and can improve performance.

Incorporating information from comments and reactions on posts can potentially improve the performance of fake news detectors. Sentiment analysis of these user comments can provide useful features to train classification models. A recent work by Cui and Lee (2019) is based on this idea—the authors have designed a model that takes into consideration user sentiments and similarity between articles based on user sentiments, along with features extracted from the image and text in the article. A modality discriminator, similar to the event discriminator in Wang et al. (2018), has been used here to drive the modality distributions for image and text to get closer to each other. Their results showed that incorporating information from user sentiment did improve the fake news detector performance.

## 3 | FACTORS BEHIND THE SPREAD OF DISINFORMATION

Social media users have a deficiency in spotting falsehood in specific emotional states, and when encountering what is consistent with their values or beliefs (Scheufele & Krause, 2019). Malicious actors use this observation and target users with the same piece of disinformation on numerous occasions. Users who receive the same content from multiple sources have higher probability of believing and spreading it (Hasher, Goldstein, & Toppino, 1977). This effective method can be further strengthened using social media bots. In this section, we first focus on emotional factors that make social media users more vulnerable to disinformation and then discuss the role of bots in boosting the effect of disinformation and methods to prevent it.

### 3.1 | Sources and publishers

Given the low cost of creating and publishing content online and the vast reach of social media platforms, several alternative media sources have emerged recently, often spreading false and/or highly biased claims. Although a large section of mainstream media is also politically/ideologically aligned toward either end of the political spectrum, these channels, having served as long-standing sources of information, do not intentionally publish false claims and disinformation. On the other hand, "alternative media" (Starbird, 2017) has seen a rise in popularity and such media sources often publish false articles, opinions disguised as facts, and even highly polarizing conspiracy theories and pseudoscience related articles. These alt-media sources publish politically motivated articles, often directly challenging the narratives as published by mainstream media. Apart from the intended target audience based on political ideology, these information sources often receive viewership from consumers of mainstream media, thus potentially fueling the reader's distrust in legitimate media sources (Haller & Holt, 2019). In this context, assessing the credibility and trustworthiness of information sources is of paramount importance. There has been some effort among research communities to determine the political bias of news articles and media sources (Hirning, Chen, & Shankar, 2017; Iyyer, Enns, Boyd-Graber, & Resnik, 2014), and also on methods to assess credibility and trustworthiness of information online (Abbasi & Liu, 2013; Moturu & Liu, 2009). Apart from the ongoing research in academia, efforts are being put in by the journalism and fact-checking community (Infogram, 2017; Media Bias/Fast Check, 2020; The Trust Project, 2017) to identify trustworthy sources and make readers aware of sources known to be publishing false articles, especially in the wake of the coronavirus pandemic.

## 3.2 | Emotional factors

Social media users have been widely affected by disinformation in recent years. For example, during the 2016 US presidential election, Guess, Nyhan, and Reifler (2018) observed that:

1. The average of 5.45 articles from fake news websites was consumed by the Americans age 18 or older.
2. Many Americans visited fake news websites to complement their hard news, not to substitute them.
3. There is a strong association between Facebook usage and fake news visits.
4. About half of the Americans who are exposed to fake news websites also visited a fact-checking websites.

The question that we aim to understand is "why do social media users believe disinformation?". Familiarity is a driver of continued influence effects, making it a risk that repeating false information, even in a fact-checking context, may increase an individual's likelihood of accepting it as true (Swire, Ecker, & Lewandowsky, 2017). Based on a study by DiFonzo and Bordia (DiFonzo & Bordia, 2007), users are more prone to propagating disinformation when the situation is uncertain, they are emotionally overwhelmed and anxious, the topic of discussion is of personal importance to them, and they do not have primary control over the situation through their actions.

### 3.2.1 | Uncertainty

Spreading fake news can be a sense-making activity in ambiguous situations and the frequency of fake news increases in uncertain situations, such as natural disasters or political events such as elections when people are unsure of the results (DiFonzo & Bordia, 2007). When a crisis happens, people first seek information from official sources. However, in the lack of such information, they form unofficial social networks to make predictions with their own judgment and fill the information gap (Rosnow & Fine, 1976). This might result in generating fake news such as a fake image of a shark swimming on the highways of Houston after Hurricane Harvey or millions of fake news posts that were shared on Facebook in the weeks leading to the US presidential election 2016. As the uncertainty increases, the reliance of firm beliefs and the unity among the users with the same ideology or in the same group reduces. Hence, users are more prone to accept new information, even false, as a compromise to resolve the uncertainty. Uncertainty can cause emotions such as anxiety and anger which will affect the spread of fake news in other ways (Marcus, 2017).

### 3.2.2 | Anxiety

Emotional pressure can play an important role in spreading fake news and can be triggered by emotions such as frustration, irritation, and anxiety. Anxiety can make people more prone to spreading unproved claims and less accurate in transmitting information (DiFonzo & Bordia, 2007). In high anxiety situations, fake news can work as a justification process to relief emotional tension (Allport & Postman, 1947). Fake news might be used as a method of expressing emotions in anxious situations that allows people to talk about their concerns and receive feedback informally; this process results in sense making and problem solving (Waddington, 2012). For example, during the devastating time of Hurricane Harvey, 2017, a fake news story accusing Black Lives Matter supporters of blocking first responders reaching the affected area was spread by more than 1 million Facebook users (Grenoble, 2017). Believing and spreading such fake news stories may help the people in disaster areas cope with the anxiety caused by delays in relief efforts (Fernandez, Alvarez, & Nixon, 2017). The recent COVID-19 Coronavirus pandemic has also brought on a wave of disinformation, false claims, and conspiracy theories to be propagated by anxious users on social media platforms, thus spreading more panic (Janosch Delcker & Scott, 2020).

### 3.2.3 | Importance or outcome-relevance

People pursue uncertainty reduction only in the areas that have personal relevance to them. For example, when a murder took place in a university campus, rumor transmission in the people from the same campus was twice the people

who were from another university campus in the same city. Due to the difficulty of measuring importance, anxiety is often used as a proxy; being anxious about a fake news story shows importance (Anthony, 1973).

#### 3.2.4 | Lack of control

Fake news represents ways of coping with uncertain and uncontrollable situations. When people do not have primary control over their situation (action-focused coping responses), they resort to secondary control strategies which are emotional responses such as predicting the worst to avoid disappointment and attributing events to chance. Two secondary control themes are explaining the meaning of events and predicting future events.

#### 3.2.5 | Belief

Users prefer information that confirms their preexisting attitudes (selective exposure), view information consistent with their preexisting beliefs as more persuasive than dissonant information (confirmation bias), and are inclined to accept information that pleases them (desirability bias; Lazer et al., 2018). Moreover, building and maintaining social relations are vital to humans, hence, to ensure their reputation as a credible source of information, they tend to share the information in which they believe. Belief is found strongly related to transmission of rumors. When it comes to political issues people tend to rationalize what they *want* to believe instead of attempting to find the truth. Hence, persuading themselves to believe what is aligned with their prior knowledge. This observation extends to the limit that appending corrections to misleading claims may worsen the situation; people who have believed the misleading claim may try to find reasons to dispute the corrections to an extent that they will believe in the misleading claim even more than before (Pennycook & Rand, 2019a, 2019b).

### 3.3 | Bots on social media

Thanks to the low cost of creating fake news sources and the software-controlled social media bots, it has never been easier to shape public opinion for political or financial reasons. Disinformation sources mimic mainstream media without obeying the same journalistic integrity (Shao et al., 2017). These sources rely on social bots to spread their content. The spread of fake news cannot be attributed to social bots only. However, curbing social bots is a promising strategy to slow down the propagation of disinformation. Actually, removing a small percentage of malicious bots can virtually eliminate the spread of low-credibility content (Shao et al., 2017).

Social bots mimic humans. They post content, interact with each other, as well as real people, and they target people that are more likely to believe disinformation. Therefore, people have a hard time distinguishing between the content posted by a human or a bot. The Botometer (Shao et al., 2017), formerly known as BotOrNot, is a machine learning tool that detects social bots on Twitter. Bots use two strategies to spread low-credibility content; first, they amplify interactions with content as soon as it gets created to speed the process of making this article go viral. Second, bots target influential users in the hope of getting them to "repost" the fabricated article to increase public exposure and thus boosting its perceived credibility (Shao et al., 2017).

In 2013, Twitter announced that about 5% of its users are fake (Elder, 2013). The Wall Street Journal reported (Koh, 2014) in March 2014 that half of the accounts created in 2014 were suspended by Twitter due to activities such as aggressive following and unfollowing behaviors which are known characteristics of bots (Lee, Eoff, & Caverlee, 2011). In 2017, Varol, Ferrara, Davis, Menczer, and Flammini (2017) reported that between 9% to 15% of users on Twitter exhibit bot behaviors. In 2018, Twitter suspended 70 million suspicious accounts (Timberg & Dwoskin, 2018) in an effort towards fighting fake news.

Political disinformation is a major activity area for bots. During the US presidential election in 2016, bots produced 1.7 billion tweets and successfully outnumbered the tweets supporting one candidate by the factor of four (Kelion & Silva, 2016). Similarly, on Facebook, millions of fake stories supporting each candidate were shared and the balance was 4 to 1 supporting the same candidate (Allcott & Gentzkow, 2017). These findings raise the question whether fake news did (Parkinson, 2016) or did not (Allcott & Gentzkow, 2017) affect the election results.

Bots, during the natural disasters of 2017, Mexico Earthquake and Hurricanes Harvey, Irma, and Maria, used disaster-related hashtags to promote political topics such as #DACA and #BlackLivesMatter (Khaund, Al-Khateeb, Tokdemir, & Agarwal, 2018). These bots shared fake news and hoaxes such as a shark swimming in a flooded highway after Hurricanes Harvey and Irma. Bots accelerate the spread of rumors and increase the number users exposed to them by 26% (Vosoughi et al., 2018). Bots also use misdirection and smoke screening to sway the public attention from specific topics (Abokhodair, Yoo, & McDonald, 2015). During the Syrian civil war in 2012, bots tweeted about political and natural crisis happening worldwide while including keywords related to the Syrian civil war to sway attention from the civil war, that is, misdirection. In smoke screening, bots talked about other events in Syria using relevant hashtags #Syria (in English and Arabic) but the content was not about the civil war.

Activities of bots in spreading disinformation on social media have been widely reflected by news agencies and vastly studied by researchers. A list of major events infiltrated by bots and reported by news media is presented in Table 1.

**TABLE 1** News articles on bots on social media during major events

| Outlet | Category | Event |
|---|---|---|
| The Guardian[a] | Elections | US Presidential Election 2016 |
| Time Magazine[b] | Elections | US Presidential Election 2016 |
| Bloomberg[c] | Elections | Italian General Election 2018 |
| The Telegraph[d] | Referendum | Brexit |
| Politico Magazine[e] | Propaganda | #ReleaseTheMemo Movement 2018 |
| The Guardian[f] | Propaganda | Russia-Ukraine Conflict 2014 |
| CNN[g] | Propaganda | Jamal Khashoggi's Death 2018 |
| The Telegraph[h] | Fake News | Stock Market—FTSE 100 Index 2018 |
| Forbes[i] | Ad Fraud | AFK13 attack |
| The New York Times[j] | Mass Shooting | Florida School Shooting 2018 |
| Fox News[k] | Mass Shooting | Texas School Shooting 2018 |
| The Telegraph[l] | Terrorist Attack | Westminster Terror Attack 2017 |
| Medium[m] | Natural Disasters | Mexico City Earthquake 2017 |

[a]https://www.theguardian.com/technology/2018/jan/19/twitter-admits-far-more-russian-bots-posted-on-election-than-it-had-disclosed.
[b]http://time.com/5286013/twitter-bots-donald-trump-votes/.
[c]https://www.bloomberg.com/news/articles/2018-02-19/now-bots-are-trying-to-help-populists-win-italy-s-election.
[d]https://www.telegraph.co.uk/technology/2018/10/17/russian-iranian-twitter-trolls-sent-10-million-tweets-fake-news/.
[e]https://www.politico.com/magazine/story/2018/02/04/trump-twitter-russians-release-the-memo-216935.
[f]https://www.cnn.com/2018/10/19/tech/twitter-suspends-spam-khashoggi-accounts-intl/index.html.
[g]https://www.cnn.com/2018/10/19/tech/twitter-suspends-spam-khashoggi-accounts-intl/index.html.
[h]https://www.telegraph.co.uk/business/2018/03/31/twitter-bots-manipulating-stock-markets-fake-news-spreads-finance/.
[i]https://www.forbes.com/sites/thomasbrewster/2016/12/20/methbot-biggest-ad-fraud-busted/#4e33e3004899.
[j]https://www.nytimes.com/2018/02/19/technology/russian-bots-school-shooting.html.
[k]https://www.foxnews.com/tech/fake-facebook-accounts-misinformation-spread-quickly-in-wake-of-santa-fe-school-shooting.
[l]https://www.telegraph.co.uk/news/2017/11/13/russian-bot-behind-false-claim-muslim-woman-ignored-victims/.
[m]https://medium.com/hci-wvu/countering-fake-news-in-natural-disasters-using-bots-and-citizen-crowds-412bbef6b489.

Researchers have studied bots on social media during numerous major events: natural disasters, man-made disasters such as civil wars and mass shootings, and political events such as Presidential Elections or referendums.

### 3.3.1 | Representative bot detection methods

Ferrara, Varol, Davis, Menczer, and Flammini (2016) proposed a taxonomy of bot detection models which divides them into three classes: (a) graph-based, (b) crowdsourcing, and (c) feature-based social bot detection methods.

*Graph-based methods*

Graph-based social bot detection models lie on the assumption that the connectivity of bots is different from human users on social media. *SybilRank* (Cao, Sirivianos, Yang, & Pregueiro, 2012) is a graph-based method proposed to efficiently detect adversary-owned bot accounts based on the links they form. The underlying assumption is that bots are mostly connected to other bots and have limited number of links to human users. Bots also show different characteristics in the communities they form. In a study on the bots that were active during the natural disasters in Khaund et al. (2018) observed that bots form more hierarchical communities with cores of bots strongly connected to each other and peripheral members who are weakly connected to the core and to each other. Moreover, human users had more communities and their communities were more tightly knit.

*Crowdsourcing methods*

Crowdsourcing social bot detection uses human annotators, expert, and hire workers, to label social media users as human or bot (Wang et al., 2013). This method is reliable and has near-zero error when the inter-annotator agreement is considered. However, it is time consuming, not cost effective, and not feasible considering millions of users on social media. Crowdsourcing and manual annotation are still being used as methods for collecting gold standard datasets for feature-based bot detection models, most of which use supervised classification.

*Feature-based methods*

Feature-based social bot detection methods are based on the observation that bots have different characteristics than human users. To use feature-based supervised bot detection models, one must identify differences among bot and human users in terms of features such as content or activity in a labeled dataset. Then, a classifier is trained on the features and labels to distinguish bots from humans in an unobserved dataset. Different classification methods can be used for this purpose such as Support Vector Machines (Morstatter, Wu, Nazer, Carley, & Liu, 2016), Random Forests (Lee et al., 2011), and Neural Networks (Kudugunta & Ferrara, 2018). We describe some common user features below:

- *Content*: The measures in this category focus on the content shared by users. Words, phrases (Varol et al., 2017), and topics (Morstatter et al., 2016) of social media posts can be a strong indicator of bot activity. Also, bots are motivated to persuade real users into visiting external sites operated by their controller, hence, share more URLs in comparison to human users (Chu, Gianvecchio, Wang, & Jajodia, 2012; Ratkiewicz et al., 2011b; Xie et al., 2008). Bots are observed to lack originality in their tweets and have large ratio of retweets/tweets (Ratkiewicz et al., 2011a).
- *Activity Patterns*: Bots tweet in a "bursty" nature (Chu et al., 2012; Lee & Kim, 2014), publishing many tweets in a short time and being inactive for a longer period of time. Bots also tend to have very regular (e.g., tweeting every 10 min) or highly irregular (randomized lapse) tweeting patterns over time (Zhang & Paxson, 2011).
- *Network Connections*: Bots connect to a large number of users hoping to receive followers back but the majority of human users do not reciprocate. Hence, bots tend to follow more users than follow them back (Chu et al., 2012).

## 4 | DETECTING DISINFORMATION

In this section, we discuss methods to detect disinformation. There are three areas that we discuss regarding detecting disinformation: (a) the users that disinformation is targeting, (b) the content of the disinformation itself, and (c) the way disinformation spreads over the network. Each one of these areas provides vital features that could be used to detect and combat disinformation. An excellent solution to the spread of disinformation would employ most, if not all, of these areas. However, most existing works had limited success to combat disinformation by focusing on one

particular area out of the ones mentioned above to build their solutions. In this section, we discuss each one of these areas in detail. Although most fake news or disinformation detection techniques are supervised, some semi-supervised (Guacho, Abdali, Shah, & Papalexakis, 2018) and unsupervised (Hosseinimotlagh & Papalexakis, 2018) techniques have also been developed. We believe that this section could help to understand the methods used to detect disinformation, and it could guide new detectors to use more than one of these areas.

## 4.1 | The role of individuals to detect disinformation

In this section, we introduce how various aspects of users who are engaged in the spreading of disinformation can be utilized to detect disinformation.

### 4.1.1 | Modeling user interactions

To combat fake news, some social media outlets turn to users to flag potential fake news articles. For instance, Facebook recently introduced tools that help users to flag articles that they deem as fake (Tschiatschek, Singla, Gomez Rodriguez, Merchant, & Krause, 2018). These tools generate crowd signals that could be used to train classifiers to flag fake news. Many fake news detectors rely on users' responses because they hold valuable information about the credibility of the news (Qian, Gong, Sharma, & Liu, 2018). Castillo, El-Haddad, Pfeffer, and Stempeck (2014) pointed out that users' responses to news articles hold valuable information that helps to understand the properties of the content itself. Also, users' stances and sentiment could assist in detecting disinformation (Qian et al., 2018).

Shu, Zhou, Wang, Zafarani, and Liu (2019) used social media's user profiles for fake news detection. They measured the sharing behavior of users that spread disinformation. Shu et al. analyzed the explicit and implicit profile features between these user groups to use those features in detecting disinformation on social media.

Tschiatschek et al. (2018) developed Detective, a Bayesian inference algorithm that detects fake news and simultaneously learns about users over time. Detective selects a subset of the news to send to an expert, such as a third-party fact-checking organization, to determine if the articles are fake or not. Over time the algorithm learns about users flagging behavior to avoid the effect of adversarial users. The experiments show that Detective is robust even if the majority of users are malicious.

Many users try to combat disinformation with the truth. Some users share links to fact checking websites that debunk false claims, or disinformation. The fake checking articles come from sites like Snopes.com, Politifact.com, or FactCheck.org. To stimulate the users to spread fact-checked content to other users, Vo and Lee (2018) propose an article recommendation model. The model provides personalized fact-checking URLs to each user based on their interests to encourage them to engage in fighting disinformation.

### 4.1.2 | Using user sentiments

User sentiment about news articles may help to detect disinformation. In a study by Vosoughi et al. (2018) show that disinformation triggers different feelings in users than real news. Disinformation sparks fear, disgust, and surprise that could be observed from user responses. On the other hand, real news stimulates anticipation, sadness, joy, and trust. Some of the existing works utilize this phenomenon to detect disinformation.

Pamungkas, Basile, and Patti (2019) argue that when users face disinformation, in this case rumors, they take different stances about the article. Some users support the rumor, while others deny it. To predict user stance, Pamungkas et al. use conversation-based and affective-based features They classify user response in these classes: support, deny, query, or comment to the rumor. This work could be used as training features to a disinformation detector.

As we discussed previously, early detection of fake news is challenging, because detectors do not have access to user responses. However, users' responses to the previous articles exist and could be used to generate responses to news articles that could enhance the detection of fake news. User responses towards previously propagated articles may hold rich information and latent user intelligence that existing works ignore (Qian et al., 2018). Qian et al. (2018) developed a detector that tackles the early fake news detection problem using only the news article's text. They developed a generative model that can be used to generate responses to new articles to assist in fake news detection.

## 4.2 | Leveraging the content to detect disinformation

In this section, we will illustrate how to leverage disinformation content in the detection task. Also, we will talk about AI-generated disinformation. We believe that AI-generated disinformation will be a very important topic for research in the near future.

As we mentioned previously, one of the challenges of detecting disinformation is detecting fake news regarding newly emerged events. Most of the existing detection strategies tend to learn event-specific features that cannot be transferred to unseen events. Therefore, they perform poorly on this challenge (Wang et al., 2018). To overcome this challenge, Wang et al. (2018) propose Event Adversarial Neural Network (EANN) that can derive event-invariant features. The event-invariant features help the detection of fake news on newly arrived events. Also, EANN can learn transferable features for unseen events.

Although most disinformation content is created manually, there is potential for AI-generated textual content in the near future. In recent years, advances in natural language understanding and natural language processing have achieved great results for challenging language tasks such as neural text summarization and machine translation. However, the same technologies used for these applications could be used to generate fake news by adversaries. We are on the verge of the era of Neural Generated Disinformation because anyone can easily and cheaply collect and process vast amounts of data (Zellers et al., 2019). In addition to that, the recent developments in next generation (Jozefowicz, Vinyals, Schuster, Shazeer, & Wu, 2016; Radford et al., 2019; Radford, Narasimhan, Salimans, & Sutskever, 2018) lead Zellers et al. (2019) to develop a model to generate fake news.

Zellers et al. (2019) developed a generative model called Grover, which is a controllable text generation model. The goal of this model to act as a threat model to study the threat of AI-generated news. This approach is inspired by threat modeling commonly used in the field of computer security. Grover can write a news article, given only a headline. For example, the user might input the title "Link Found Between Vaccines and Autism" as a headline for a new fake news article. Grover then generates the body of the article and rewrites the headline of the article in a more appropriate way. Grover can identify news generated by itself with around 90% accuracy. AI-Generated content could possibly lead to other threats. For instance, AI could be used by an adversary to generate comments to news articles or even the articles themselves (Zellers et al., 2019).

Media and journalism companies are coming up with smartphone apps and websites that provide very short summaries of news articles, to satisfy readers who want to consume as much information possible in a short time. In this regard, neural text summarization seems to be a promising way of automating the news article summarization task. With the gradual improvement of such summarization models, an interesting new challenge would be to perform fake news detection on such neural generated summaries instead of on the actual article. Esmaeilzadeh, Peh, and Xu (2019) have performed fake news detection on summaries generated from news articles using abstractive summarization models, and compared the detection performance when using the entire news text and headline text as input. For the dataset and text summarization model used in Esmaeilzadeh et al. (2019), the model trained on the neural generated summaries performed best when it came to detection accuracy, hinting that the summarization step acts as some form of a feature generator.

Apart from the ML-based methods discussed above, significant efforts have been put in by the fact-checking, data management, and database community (Lakshmanan, Simpson, & Thirumuruganathan, 2019) toward disinformation detection, leveraging structured data, and knowledge graphs (Ciampaglia et al., 2015). Automatic fact-checking using *knowledge graphs* to check the veracity of claims is an active area of ongoing research and when done successfully can potentially replace the tedious task of manual fact-checking done by experts and journalists. Such techniques include path analysis (Shi & Weninger, 2016a) and link prediction (Shi & Weninger, 2016b) in knowledge graphs among others.

Attempts at disinformation detection via checking the credibility of claims made online has been done via methods in the domain of *Truth Discovery and Fusion*. Some of these methods use multiple online sources, both supporting and refuting, to check the veracity of claims and identify the truth (Jin, Cao, Zhang, & Luo, 2016; Yin, Han, & Philip, 2008). Given the evolving nature of online social media content along with the problem of early-detection of disinformation, Zhang et al. have proposed a dynamic truth discovery scheme (Zhang, Wang, & Zhang, 2017) to deal with noisy and incomplete data available online.

## 4.3 | Exploiting networks for detecting disinformation

In this section, we illustrate how to detect disinformation with the signals extracted from different types of networks.

### 4.3.1 | Detection of fake content via propagation networks

Malicious agents, who intend to spread fake fabricated news on social media, want rapid dissemination of the post/article, to reach as many people as possible quickly. Often bots and bot-like accounts are used to propagate the news faster. Keeping this in mind, the user profiles of users sharing the news, and also the propagation pattern can provide clues to determine the veracity of the article. One relatively recent work that achieved promising results using this approach is that of Liu and Wu (2018), where the authors have used the concept of propagation path of news articles in the detection of fake news. In this work, propagation path of an article is a sequence of tuples, each tuple consisting of a timestamp $t$ and a vector characterizing the user who shared the article at that timestamp. The propagation paths for the news articles, are fed separately into RNN and CNN units, and after subsequent pooling and concatenation operations, a classifier is trained on the learned representations to differentiate between fake and real articles. A similar approach using propagation structures or propagation trees, and features related to the user, original message and "reposts", has been experimented with using false rumor versus real data from Weibo (Wu, Yang, & Zhu, 2015).

### 4.3.2 | Advanced graph mining for disinformation detection

Recently, much efforts have been devoted to graph mining-based disinformation detection due to their superior capabilities. Graph neural networks (GNNs), a family of neural models for learning latent node representations in a graph, have been widely used in different graph learning tasks and achieved remarkable success (Bian et al., 2020; Ding, Li, Bhanushali, & Liu, 2019; Ding, Li, Li, Liu, & Liu, 2019; Kipf & Welling, 2016; Monti, Frasca, Eynard, Mannion, & Bronstein, 2019). Due to the superior modeling power, researchers have proposed to leverage GNNs for solving the disinformation detection problem. As one of the first endeavors, Monti et al. (2019) has presented a geometric deep learning approach for fake news detection on Twitter social network. The proposed method allows integrating heterogeneous data pertaining to the user profile and activity, social network structure, news spreading patterns, and content. The proposed model achieves promising and robust results on the large-scale dataset, pointing to the great potential of GNN-based methods for fake news detection. Later on, Graph Convolutional Networks based Source Identification (GCNSI) has been proposed to locate multiple rumor sources without prior knowledge of underlying propagation model. By adopting spectral-domain convolution, GCNSI learns node representation by utilizing its multi-order neighbors information such that the prediction precision on the sources is improved. Moreover, Bian et al. (2020) argue that the existing methods only take into account the patterns of deep propagation but ignore the structures of wide dispersion in disinformation (rumor) detection. They propose a novel bi-directional graph model, named Bi-Directional Graph Convolutional Networks (Bi-GCN), to explore both characteristics by operating on both top-down and bottom-up propagation of rumors. Its inherent GCN model gives the proposed method Bi-GCN the ability of processing graph/tree structures and learning higher-level representations more conducive to disinformation (rumor) detection.

## 5 | MITIGATING DISINFORMATION VIA EDUCATION, RESEARCH, AND COLLABORATION

In recent years, especially after the 2016 US election, disinformation has become a worry to researchers from different fields and government officials. We noticed a surge in research related to identifying, understanding, and combating disinformation. Also, we see that many governments and educational organizations around the world are trying to combat disinformation. In this section, we discuss the research fields that are concerned about disinformation, and we list the strategies they give to combat disinformation. In addition to that we talk about efforts related to combating disinformation through education.

### 5.1 | Mitigating disinformation via education

In an alarming survey report by the Stanford History Education Group (Breakstone et al., 2019), over 3,400 high school students in the United States were evaluated on their ability to differentiate "fact from fiction" and how well they were able to judge the credibility of information sources on online social media. According to the report, students were

assigned six tasks that evaluated how they consume news on social media, and majority of the students failed at the tasks they were assigned, thus proving the need for administering digital media evaluation skills to students early on in their life.

Recognizing the alarming prevalence of fake news and disinformation, and the potential threat to consumers of such false information, many countries, government, and private bodies have been investing in steps to educate the general public about disinformation and how to spot fake news. Many schools across all 50 states in the US (Chen, 2018; Timsit, 2019; Tugend, 2020) have modified their curricula to add courses to educate students to be critical to the news they see on social media. *News Literacy Project*[1]—a US education non-profit—is responsible for designing the curriculum and offering nonpartisan, independent programs that teach students how to know what to trust in the digital age. Recently, Google has also launched media literacy activities to teach kids how to be aware of the information they see online, and how to identify fake posts and URLs (Mascott, 2019). They have also partnered with YMCA and National PTA to host workshops for children and parents.

Similar approaches to educate citizens are being taken across the globe. Finland, which ranks the highest in terms of media literacy among European Nations (Lessenski, 2018), have schools and colleges teaching students how to evaluate the authenticity of articles on social media before they "like" or "share" them.

## 5.2 | Mitigating disinformation via research and collaboration

Many researchers from different fields are trying to understand disinformation to mitigate its effect on society. For instance, cognitive scientists study the problem of disinformation through studying why do people believe disinformation and what type of people are more likely to believe it (Bronstein, Pennycook, Bear, Rand, & Cannon, 2019). Journalists are trying to re-evaluate the reporting of the news and how "elites" should discuss disinformation in the media (Van Duyn & Collier, 2019). Computer scientists are applying data mining and machine learning to detect and deal with disinformation (Section 4). Here, we talk about these efforts, and we list some of the strategies that researchers are suggesting to mitigate the spread of disinformation.

### 5.2.1 | Computational approaches

Given the prevalence of disinformation, websites, and social media platforms have been focusing a lot of their computational resources on mitigation methods (Shu, Bernard, & Liu, 2018), some of which are briefly explained as follows.

*Source identification*: Identifying malicious sources involved in the creation and spread of fake news and disinformation. An important step in this process is the task of information provenance—identifying the sources given the diffusion network of the news article (Barbier, Feng, Gundecha, & Liu, 2013), using provenance paths. Influential social media users having a large online follower-base may also be responsible for the rapid spread of disinformation. Identifying these "opinion leaders" propagating disinformation and terminating their accounts or even slowing down their reach or may slow down the spread of disinformation.

*Network intervention*: Various algorithmic techniques have been developed to slow down and limit the spread of disinformation. Most of the techniques are based on the concept of diffusion of information on social network of users, following the dynamics of the Independent Cascade Model or Linear Threshold Model. Researchers tackling this issue have proposed techniques like Influence Minimization or Influence Limitation. Some of these approaches focus on designing cascades of counter campaigns such that the disinformation cascade slows down (Budak, Agrawal, & Abbadi, 2011; Tong, Du, & Wu, 2018), while some others focus on finding the smallest set of nodes to remove, such that the influence is minimized (Pham, Phu, Hoang, Pei, & Thai, 2019).

*Content Flagging*: Social media platforms often provide users the option to "flag" a post as spreading false information, and once a post gets sufficient number of these "flags," it is often fact-checked, and if found to be disinformation, it is subsequently removed from the platform. Given the trade-off between the number of user flags and the potential damage caused by increased exposure to disinformation, Kim, Tabibian, Oh, Schölkopf, and Gomez-Rodriguez (2018) proposed a scalable algorithm to effectively select articles for fact-checking and schedule the fact-checking process.

### 5.2.2 | Cognitive science

Cognitive science is defined as the scientific study of the human mind to understand how human knowledge used, processed, and acquired.[2] Cognitive scientists are concerned about disinformation, and they are applying their knowledge to find how and why people believe disinformation. Bronstein et al. (2019) studied the reasons that might lead people to believe in fake news. They found that the main factors behind believing disinformation are Delusionality, Dogmatism, Religious Fundamentalism, and Reduced Analytic Thinking. Bronstein et al. concluded that people who uphold delusion-like beliefs, dogmatic individuals, or religious fundamentalists are more likely to believe disinformation. They argue that the reason behind this phenomenon is that these people are less likely to engage in open-minded thinking and analytic thinking, which makes them vulnerable to disinformation. Pennycook and Rand (2019a, 2019b) found that believing disinformation is more related to lazy thinking or lack of it more than partisan bias.

The lack of thinking and other factors that lead to believing disinformation is a solvable problem. Keersmaecker and Roets (2017) demonstrate that, generally, people adjust their attitudes toward disinformation once they are faced with the truth. However, the level of cognitive ability influences the degree to which people change their assessment of the news. Cognitive scientist believes that leveraging interventions that improve analytic and actively open-minded thinking might lessen the belief in disinformation (Bronstein et al., 2019; Pennycook & Rand, 2019a, 2019b). Actively open-minded thinking "involves the search for alternative explanations and the use of evidence to revise beliefs." Furthermore, analytic thinking is "the disposition to initiate deliberate thought processes to reflect on intuitions and gut feelings" (Bronstein et al., 2019).

### 5.2.3 | Journalism and political science

Following the 2016 US election, many political scientists and journalists studied the effect of disinformation on social media on society. Also, they studied disinformation to find the people that are more likely to be victims of fake news to develop strategies to mitigate the problem of disinformation. Bovet and Makse (2019) found that confirmation bias and social influence is the main reason for echo chambers where users with similar beliefs share disinformation about a specific topic. However, users who are exposed to disinformation is a small fraction of the general public (Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019). Those users tend to be conservative-leaning, older, and highly engaged with political news (Grinberg et al., 2019). Those users tend to consume disinformation about political news (Grinberg et al., 2019) or science (Scheufele & Krause, 2019).

Journalists and political scientists wanted to study the reasons behind the distrust of traditional media. Van Duyn and Collier (2019) wanted to study the effect of elite discourse on disinformation and how does it affect people's trust of news organization. They found that people exposed to the elite discussion about disinformation tend to have lower levels of trust in traditional media, and they have a hard time identifying real information. Also, attacking real news organizations, and branding them as fake news sites lead to the rise of disinformation on social media (Tandoc Jr, 2019).

Journalists and political scientists suggest that users should consume news from verified accounts such as well-known journalists or official accounts such as government accounts (Bovet & Makse, 2019). In addition to that, disinformation experts should be careful about the way they discuss disinformation to avoid causing distrust in the media in general (Van Duyn & Collier, 2019). Politicians should also stop attacking the real news media and label them as fake news.

## 6 | CONCLUSION AND FUTURE WORK

In this article, we gave an overview of disinformation from different aspects. We defined disinformation, and listed some of the forms it takes, such as fake news, rumors, hoaxes, and so on. We also talked about the history of disinformation and how the Internet and social media affect the spread and subsequent consumption of disinformation. We discuss the challenges in detecting fake news and then proceed to talk about the different forms of fabricated content that may exist on social media, that is, text, image, video, and so on, and discuss the notable detection techniques devised by researchers for detecting the different forms of fabricated content. We also talked about the ongoing efforts to inform people about the presence of disinformation, in terms of educational programs, and also gave a brief

perspective into disinformation through the lenses of cognitive science and political science. The main goal of our work is to provide the reader with a comprehensive view of the current scenario in the research field of disinformation detection. Given the ever-growing relevance of disinformation on social media, and the complementary need for better detection methods and models, we hope our work would provide future researchers an idea of the advances made so far, and hence guide them towards further improvement. In this regard, we also list some of the datasets and tools used most widely in the context of disinformation detection.

There are several interesting and promising directions to explore. First, threat modeling of disinformation is an exciting topic to study. Threat modeling is a widely used technique in the field of computer security to identify and combat the threat. We believe that threat modeling is a very interesting topic that was not studied to its full potential. Disinformation generators could be developed as threat models to improve our understanding of disinformation and how it spreads in society. Building threat models may help to build better disinformation detectors. To the best of our knowledge, there is no work on disinformation threat modeling other than Zellers et al. (2019). However, Zellers et al. (2019) work is limited to only textual data. As we explained previously, real-world disinformation comes in many forms, such as text, images, and videos. To build better threats models, these models need to be able to generate disinformation in as many forms as possible.

Now that generating deepfakes and fake videos in general has become extremely easy, we feel there is a need for more work to be done in the detection of such fabricated videos. There is a great need for video threat modeling. One possible way to create these models is by using stock video databases available online. This method is inexpensive and quick to overwhelm detectors. Moreover, very few labeled deepfake datasets are available publicly for research. Work could be done in generating and creating more annotated deepfake datasets.

The way that social media users react to novel news content is an excellent way to detect disinformation. People tend to respond differently to disinformation and real information. However, some people are more likely to get fooled by disinformation than others. The reason behind this phenomenon might be confirmation bias or some other cognitive or psychological reasons, and more research could be done to study disinformation from a psychological point of view. People who get fooled by disinformation easily could be used as an indicator of disinformation. Also, those people tend to form echo chambers, and they share disinformation among themselves. Therefore, we believe that tackling the problem of echo chambers could possibly help early detection.

As we mentioned previously, the problem of early detection of fake news on social media platforms is an unsolved one and much improvements could be done in this regard. Furthermore, the problem of fake news is extremely topic-dependent. Hence, in the case of ML-based methods, models trained using data belong to a certain domain or set of events may fail to perform satisfactorily in a different context. To tackle this challenge, along with the challenge of scarcity of available data when it comes to novel fake news, transfer learning-based techniques could be used. Newly available data could be used to further tune a model that has already learned the linguistic structures of fake versus real news.

## CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

## AUTHOR CONTRIBUTIONS

**Amrita Bhattacharjee:** Investigation; resources; writing-original draft; writing-review and editing. **Faisal Alatawi:** Investigation; resources; writing-original draft; writing-review and editing. **Tahora H. Nazer:** Writing-original draft; writing-review and editing. **Kaize Ding:** Writing-original draft; writing-review and editing. **Mansooreh Karami:** Writing-original draft; writing-review and editing. **Huan Liu:** Conceptualization; supervision; writing-original draft; writing-review and editing.

## ORCID

*Kai Shu* https://orcid.org/0000-0002-6043-1764
*Amrita Bhattacharjee* https://orcid.org/0000-0001-6117-6382
*Huan Liu* https://orcid.org/0000-0002-3264-7904

## ENDNOTES

[1] https://newslit.org/.

[2] https://bcs.mit.edu/research/cognitive-science.

[3] https://github.com/MKLab-ITI/image-verification-corpus.

[4] https://figshare.com/articles/PHEME_dataset_for_Rumour_Detection_and_Veracity_Classification/6392078.

[5] https://github.com/compsocial/CREDBANK-data.

[6] https://www.cs.ucsb.edu/~William/data/liar_dataset.zip.

[7] https://figshare.com/articles/Twitter_Death_Hoaxes_dataset/5688811.

[8] https://doi.org/10.5281/zenodo.2607278.

[9] https://github.com/KaiDMML/FakeNewsNet.

[10] https://bigbird.dev/.

[11] https://github.com/MKLab-ITI/computational-verification.

[12] https://github.com/gabll/some-like-it-hoax/tree/master/dataset.

## RELATED WIREs ARTICLE

Credibility in social media: Opinions, news, and health information—A survey

## FURTHER READING

Allem, J.-P., Ferrara, E., Uppu, S. P., Cruz, T. B., & Unger, J. B. (2017). E-cigarette surveillance with social media data: Social bots, emerging topics, and trends. *JMIR Public Health and Surveillance*, *3*(4), e98.

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. presidential election online discussion. *First Monday*, http://dx.doi.org/10.5210/fm.v21i11.7090.

Broniatowski, D. A., Jamison, A. M., Qi, S. H., AlKulaib, L., Chen, T., Benton, A., ... Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, *108*(10), 1378–1384.

Jin, Z., Cao, J., Guo, H., Zhang, Y., Wang, Y., & Luo, J. (2017). *Detection and analysis of 2016 us presidential election related rumors on twitter*. In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (pp. 14–24). Springer.

Kitzie, V. L., Karami, A., & Mohammadi, E. (2018). "Life Never Matters in the Democrats Mind": Examining Strategies of Retweeted Social Bots During a Mass Shooting Event. *arXiv:1808.09325*. Retrieved from http://arxiv.org/pdf/1808.09325v1.

Littman, J., Kerchner, D., He, Y., Tan, Y., & Zeljak, C. (2017). Collecting social media data from the Sina Weibo Api. *Journal of East Asian Libraries*, *2017*(165), 12.

Nied, A. C., Stewart, L., Spiro, E., & Starbird, K. (2017). *Alternative narratives of crisis events*: *Communities and social botnets engaged on social media*. In Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 263–266). ACM.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), 4787.

Shao, C., Hui, P. M., Wang, L., Jiang, X., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018). Anatomy of an online misinformation network. *PLoS One*, *13*(4), 1–23.

Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots sustain and inflate striking opposition in online social systems. *arxiv* 1–10. Retrieved from http://arxiv.org/abs/1802.07292

Thomas, K., Grier, C., & Paxson, V. (2012). *Adapting social spam infrastructure for political censorship*. In Conference on Large-Scale Exploits and Emergent Threats. USENIX.

## REFERENCES

Abbasi, M.-A., & Liu, H. (2013). Measuring user credibility in social media. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 441–448). Berlin: Springer.

Abokhodair, N., Yoo, D., & McDonald, D. W. (2015). Dissecting a social botnet: Growth, content and influence in twitter. In *Proceedings of the 18th ACM conference on Computer Supported Cooperative Work & Social Computing* (pp. 839–851). New York, NY: ACM Publications.

Allcott, H., & Gentzkow, M. (2017). *Social media and fake news in the 2016 election*. Cambridge, MA: National Bureau of Economic Research.

Allport, G. W., & Postman, L. (1947). *The psychology of rumor*. Oxford, England: Henry Holt.

Anthony, S. (1973). Anxiety and rumor. *The Journal of Social Psychology*, *89*, 91–98.

Arun, C. (2019). On WhatsApp, Rumours, and Lynchings. *Economic & Political Weekly*, *54*(6), 30–35.

Barbier, G., Feng, Z., Gundecha, P., & Liu, H. (2013). Provenance Data in Social Media. *Synthesis Lectures on Data Mining and Knowledge Discovery*, *4*(1), 1–84.

Bettencourt, L. M. A., Cintrón-Arias, A., Kaiser, D. I., & Castillo-Chávez, C. (2006). The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. *Physica A: Statistical Mechanics and its Applications*, *364*, 513–536.

Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., & Huang, J. (2020). Rumor detection on social media with bi-directional graph convolutional networks. *arXiv Preprint arXiv*, *2020*, 2001.06362.

Boghardt, T. (2009). Soviet bloc intelligence and its AIDS disinformation campaign. *Studies in Intelligence*, *53*(4), 1–24.

Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., & Kompatsiaris, Y. (2018). Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, *7*(1), 71–86 Springer.

Bovet, A., & Makse, H. A. (2019). Influence of fake news in twitter during the 2016 US presidential election. *Nature Communications*, *10* (1), 1–14.

Breakstone, J., Smith, M., Wineburg, S., Rapaport, A., Carle, J., Garland, M., & Saavedra, A. (2019). *Student's civic online reasoning: A national portrait*. Stanford, CA: Stanford History Education Group & Gibson Consulting.

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, *8*(1), 108–117.

Budak, C., Agrawal, D., & Abbadi, A. E. (2011). *Limiting the spread of misinformation in social networks*. In Proceedings of the 20th International Conference on World Wide Web - WWW 11. ACM Press. Retrieved from https://doi.org/10.1145%2F1963405.1963499

Cao, Q., Sirivianos, M., Yang, X., & Pregueiro, T. (2012). *Aiding the detection of fake accounts in large scale social online services*. In Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation (p. 15). USENIX Association.

Castillo, C., El-Haddad, M., Pfeffer, J., & Stempeck, M. (2014). *Characterizing the life cycle of online news stories using social media reactions*. In Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (pp. 211–223).

Chen, S. (2018). Schools around the world are now teaching kids to spot fake news. *Quartz*. Retrieved from https://qz.com/1175155/a-special-class-how-to-teach-kids-to-spot-fake-news/.

Chollet, F. (2017). *Xception: Deep learning with depthwise separable convolutions*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1251–1258).

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811–824 IEEE.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. *PLoS One*, *10*(6), e0128193.

Cui, L., & Lee, S. W. D. (2019). SAME: Sentiment-aware multi-modal embedding for detecting fake news.

De Maeyer, D. (1997). Internet's information highway potential. *Internet Research*.

Department of State. (1981). Forgery, disinformation and political operation. *Department of State Bulletin*, *81*(2056), 52–55.

DiFonzo, N., & Bordia, P. (2007). *Rumor psychology: Social and organizational approaches*. Washington, DC: American Psychological Association.

Ding, K., Li, J., Bhanushali, R., & Liu, H. (2019). *Deep anomaly detection on attributed networks*. In Proceedings of the 2019 SIAM International Conference on Data Mining, SIAM. pp. 594–602.

Ding, K., Li, Y., Li, J., Liu, C., & Liu, H. (2019). Graph neural networks with high-order feature interactions. *arXiv Preprint arXiv*, *2019*, 1908.07110.

Elder, J. (2013, November). Inside a twitter robot factory; fake activity, often bought for publicity purposes, influences trending topics. *Wall Street Journal (Online)*. Retrieved from https://www.wsj.com/articles/bogus-accounts-dog-twitter-1385335134.

Esmaeilzadeh, S., Peh, G. X., & Xu, A. (2019). Neural abstractive text summarization and fake news detection. *arXiv preprint arXiv*: *1904.00788*.

Fernandez, M., Alvarez, L., & Nixon, R. (2017, October 22). Still Waiting for FEMA in Texas and Florida After Hurricanes. *The New York Times*. Retrieved from https://www.nytimes.com/2017/10/22/us/fema-texas-florida-delays-.html.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104 ACM.

Fetzer, J. H. (2004). Disinformation: The use of false information. *Minds and Machines*, *14*(2), 231–240 Springer.

George, S. (2019, Januray 13). 'Deepfakes' called new election threat, with no easy fix. *AP News*.

Goldenberg, J., Libai, B., & Muller, E. (2001a). Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. *Academy of Marketing Science Review*, *9*(3), 1–18.

Goldenberg, J., Libai, B., & Muller, E. (2001b). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, *12*(3), 211–223.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ..., Benjio, Y. (2014). *Generative adversarial nets*. In Proceedings of the International Conference on Advances in Neural Information Processing Systems 27, Montreal, Quebec, Canada, pp. 2672–2680.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, *83*(6), 1420–1443 University of Chicago Press.

Grenoble, R. (2017, July 9). Hurricane Harvey is just the latest in Facebook's fake news problem. Huffington Post. Retrieved from https://www.huffingtonpost.com/entry/facebook-hurricane-harvey-fake-news_us_59b17900e4b0354e441021fb.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on twitter during the 2016 US presidential election. *Science*, *363*(6425), 374–378.

Guacho, G. B., Abdali, S., Shah, N., & Papalexakis, E. E. (2018). *Semi-supervised Content-Based Detection of Misinformation via Tensor Embeddings*. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Retrieved from https://doi.org/10.1109%2Fasonam.2018.8508241.

Güera, D., & Delp, E. J. (2018). *Deepfake video detection using recurrent neural networks*. In 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6.

Guess, A., Nyhan, B., & Reifler, J. (2018). Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. Retrieved from https://www.dartmouth.edu/nyhan/fake-news-2016.pdf.

Haller, A., & Holt, K. (2019). Paradoxical populism: How PEGIDA relates to mainstream and alternative media. *Information, Communication & Society*, *22*(12), 1665–1680.

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107–112.

Hernandez, J. C., Hernandez, C. J., Sierra, J. M., & Ribagorda, A. (2002). *A first step towards automatic hoax detection*. In Proceedings of the 36th Annual 2002 International Carnahan Conference on Security Technology. Retrieved from https://doi.org/10.1109%2Fccst.2002.1049234.

Hernon, P. (1995). Disinformation and misinformation through the internet: Findings of an exploratory study. *Government Information Quarterly*, *12*(2), 133–139.

Hirning, N. P., Chen, A., & Shankar, S. (2017). Detecting and identifying bias-heavy sentences in news articles. Technical report, Stanford University, Stanford, CA.

Hosseinimotlagh, S., & Papalexakis, E. E. (2018). *Unsupervised content-based identification of fake news articles with tensor decomposition ensembles*. Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2).

Infogram(2017, November 9). PolitiFact's fake news almanac. Retrieved from https://infogram.com/politifacts-fake-news-almanac-1gew2vjdxl912nj.

Iyyer, M., Enns, P., Boyd-Graber, J., & Resnik, P. (2014). *Political ideology detection using recursive neural networks*. In Proceedings of the 52nd annual meeting of the Association for Computational Linguistics. Volume 1: Long Papers. pp. 1113–1122.

Janosch Delcker, Z. W., & Scott, M. (2020). The coronavirus fake news pandemic sweeping WhatsApp. *Politico*. Retrieved from https://www.politico.com/news/2020/03/16/coronavirus-fake-news-pandemic-133447

Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013). *Epidemiological modeling of news and rumors on twitter*. In Proceedings of the 7th Workshop on Social Network Mining and Analysis (p. 8). ACM.

Jin, Z., Cao, J., Zhang, Y., & Luo, J. (2016). *News verification by exploiting conflicting social viewpoints in microblogs*. In Thirtieth AAAI Conference on Artificial Intelligence.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. *arXiv Preprint arXiv*: 1602.02410.

Keersmaecker, J. D., & Roets, A. (2017). 'Fake news': Incorrect but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence*, *65*, 107–110. https://doi.org/10.1016%2Fj.intell.2017.10.005

Kelion, L., & Silva, S. (2016). Pro-Clinton bots 'fought back but outnumbered in second debate'. *BBC News*. Retrieved from http://www.bbc.com/news/technology-37703565.

Khaund, T., Al-Khateeb, S., Tokdemir, S., & Agarwal, N. (2018). *Analyzing social bots and their coordination during natural disasters*. In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (pp. 207–212). Springer.

Kim, J., Tabibian, B., Oh, A., Schölkopf, B., & Gomez-Rodriguez, M. (2018). *Leveraging the crowd to detect and reduce the spread of fake news and misinformation*. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (pp. 324–332).

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kochkina, E., Liakata, M., & Zubiaga, A. (2018). PHEME dataset for Rumour Detection and Veracity Classification. Retrieved from https://figshare.com/articles/PHEME_dataset_for_Rumour_Detection_and_Veracity_Classification/6392078.

Koh, Y. (2014, March). Only 11% of New Twitter Users in 2012 Are Still Tweeting. *Dow Jones Institutional News*. Retrieved from https://blogs.wsj.com/digits/2014/03/21/new-report-spotlights-twitters-retention-problem/.

Kudugunta, S., & Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences*, *467*, 312–322.

Kumar, S., West, R., & Leskovec, J. (2016). *Disinformation on the web*. In Proceedings of the 25th International Conference on World Wide Web, WWW '16. ACM Press. Retrieved from https://doi.org/10.1145%2F2872427.2883085.

Lakshmanan, L. V. S., Simpson, M., & Thirumuruganathan, S. (2019). Combating fake news: A data management and mining perspective. *Proceedings of the VLDB Endowment*, *12*(12), 1990–1993.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Lee, D. (2018, February 3). Deepfakes porn has serious consequences. *BBC News*.

Lee, K., Eoff, B. D., & Caverlee, J. (2011). Seven months with the devils: A long-term study of content polluters on twitter. In *ICWSM* (pp. 185–192). AAAI. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2780

Lee, S., & Kim, J. (2014). Early filtering of ephemeral malicious accounts on twitter. *Computer Communications*, *54*, 48–57.

Lessenski, M. (2018). *Common sense wanted: Resilience to 'post-truth' and its predictors in the new media literacy index 2018*. Manhattan, NY: Open Society Institute.

Li, Y., Chang, M.-C., & Lyu, S. (2018). In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. *arXiv Preprint arXiv*: 1806.02877.

Liu, Y., & Wu, Y.-F. B. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In Thirty-second AAAI Conference on Artificial Intelligence.

Manning, M. J., Manning, M., & Romerstein, H. (2004). *Historical dictionary of American propaganda*. West Port, CT: Greenwood Publishing Group.

Marcus, G. (2017). How affective intelligence can help us understand politics. *Emotion Researcher*. Retrieved from https://emotionresearcher.com/how-affective-intelligence-theory-can-help-us-understand-politics/

Marra, F., Gragnaniello, D., Cozzolino, D., & Verdoliva, L. (2018). Detection of GAN-generated fake images over social networks. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 384–389. IEEE.

Mascott, A. (2019). Helping kids learn to evaluate what they see online. Retrieved from https://blog.google/technology/families/be-internet-awesome-media-literacy/.

Media Bias/Fast Check(2020). Search and Learn the Bias of News Media. Retrieved from https://mediabiasfactcheck.com/.

Mitra, T., & Gilbert, E. (2015). Credbank: A large-scale social media corpus with associated credibility annotations. In Ninth International AAAI Conference on Web and Social Media. Retrieved from https://github.com/compsocial/CREDBANK-data.

Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake News Detection on Social Media using Geometric Deep Learning. *arXiv preprint arXiv*:1902.06673.

Morstatter, F., Wu, L., Nazer, T. H., Carley, K. M., & Liu, H. (2016). A new approach to bot detection: striking the balance between precision and recall. In 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), (pp. 533–540). IEEE.

Moturu, S. T., & Liu, H. (2009). *Evaluating the trustworthiness of Wikipedia articles through quality and credibility*. In Proceedings of the 5th International Symposium on Wikis and Open Collaboration (pp. 1–2).

Nami Sumida, M. W., & Mitchell, A. (2019). The role of social media in news. *Pew Research Center - Journalism and* Media. Retrieved from https://www.journalism.org/2019/04/23/the-role-of-social-media-in-news/.

Nataraj, L., Mohammed, T. M., Manjunath, B. S., Chandrasekaran, S., Flenner, A., Bappy, J. H., & Roy-Chowdhury, A. K. (2019). Detecting GAN generated fake images using co-occurrence matrices. *arXiv preprint arXiv:1903.06836*.

Nguyen, N. P., Yan, G., Thai, M. T., & Eidenbenz, S. (2012). *Containment of misinformation spread in online social networks*. In Proceedings of the 4th Annual ACM Web Science Conference (pp. 213–222).

Odena, A., Dumoulin, V., & Olah, C. (2016). Deconvolution and checkerboard artifacts. *Distill*, *1*(10), e3.

Pamungkas, E. W., Basile, V., & Patti, V. (2019). Stance classification for rumour analysis in Twitter: Exploiting affective information and conversation structure. *arXiv preprint arXiv:1901.01911*.

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. London: Penguin.

Parkinson, H. J. (2016). Click and elect: How fake news helped Donald Trump win a real election. *The Guardian*. Retrieved from https://goo.gl/DJiWNd

Parth, M. N., Bengali, S.. (2018, May 30). Rumors of child-kidnapping gangs and other WhatsApp hoaxes are getting people killed in India. *Los Angeles Times*.

Patrick J. McDonnell, C. S. (2018, September 21). When fake news kills: Lynchings in Mexico are linked to viral child-kidnap rumors. *Los Angeles Times*.

Pennycook, G., & Rand, D. G. (2019a, January 19). Why do people fall for fake news? *The New York Times*. Retrieved from https://www.nytimes.com/2019/01/19/opinion/sunday/fake-news.html.

Pennycook, G., & Rand, D. G. (2019b). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Pham, C. V., Phu, Q. V., Hoang, H. X., Pei, J., & Thai, M. T. (2019). Minimum budget for misinformation blocking in online social networks. *Journal of Combinatorial Optimization*, *38*(4), 1101–1127.

Pratiwi, I. Y. R., Asmara, R. A., & Rahutomo, F. (2017). Study of hoax news detection using naïve bayes classifier in Indonesian language. In *2017 11th International Conference on Information & Communication Technology and System (ICTS)*. IEEE. Retrieved from https://doi.org/10.1109%2Ficts.2017.8265649.

Qian, F., Gong, C., Sharma, K., & Liu, Y. (2018). *Neural user response generator: Fake news detection with collective user intelligence*. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. pp. 3834–3840.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., & Menczer, F. (2011b). Truthy: Mapping the spread of astroturf in microblog streams. In *World Wide Web Companion* (pp. 249–252). ACM.

Ratkiewicz, J., Conover, M., Meiss, M. R., Gonçalves, B., Flammini, A., & Menczer, F. (2011a). Detecting and tracking political abuse in social media. *ICWSM*, *11*, 297–304.

Rosnow, R. L. (1991). Inside rumor: A personal journey. *American Psychologist*, *46*(5), 484–496.

Rosnow, R. L., & Fine, G. A. (1976). *Rumor and gossip: The social psychology of hearsay*. Amsterdam: Elsevier.

Salem, F. K. A., Al Feel, R., Elbassuoni, S., Jaber, M., & Farah, M. (2019). FA-KES: A fake news dataset around the Syrian war. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. *13*, pp. 573–582).

Santoso, I., Yohansen, I., Nealson, Warnars, H. L. H. S., & Hashimoto, K. (2017). *Early investigation of proposed hoax detection for decreasing hoax in social media*. In 2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom). IEEE. Retrieved from https://doi.org/10.1109%2Fcyberneticscom.2017.8311705

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, *116*(16), 7662–7669.

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (2017). The spread of fake news by social bots. *arXiv preprint arXiv*: *1707.07592*, *96*, 104. ArXiv e-prints.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(3), 1–42.

Shearer, E., & Matsa, K. E. (2018). News use across social media platforms 2018. *Pew Research Center - Journalism and Media*. Retrieved from https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/.

Shi, B., & Weninger, T. (2016a). Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems*, *104*, 123–133.

Shi, B., & Weninger, T. (2016b). Fact checking in heterogeneous information networks. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 101–102).

Shu, K., Bernard, H. R., & Liu, H. (2018). Studying fake news via network analysis: Detection and mitigation. In *Lecture notes in social networks* (pp. 43–65). Cham: Springer International Publishing Retrieved from https://doi.org/10.1007%2F978-3-319-94105-9_3

Shu, K., & Liu, H. (2019). Detecting fake news on social media. *Synthesis Lectures on Data Mining and Knowledge Discovery*, *11*(3), 1–129.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2018). FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *arXiv preprint arXiv:1809.01286*.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Shu, K., Zhou, X., Wang, S., Zafarani, R., & Liu, H. (2019). *The role of user profiles for fake news detection*. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 436–439).

Starbird, K. (2017). *Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter*. In Eleventh International AAAI Conference on Web and Social Media.

Sunstein, C. R., & Vermeule, A. (2009). Conspiracy theories: Causes and cures. *Journal of Political Philosophy*, *17*(2), 202–227.

Swire, B., Ecker, U. K. H., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology-Learning Memory and Cognition*, *43*(12), 1948–1961.

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L.. (2017). Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.

Tandoc, E. C., Jr. (2019). The facts of fake news: A research review. *Sociology Compass*, *13*(9), e12724.

Tariq, S., Lee, S., Kim, H., Shin, Y., & Woo, S. S. (2018). *Detecting both machine and human created fake face images in the wild*. In Proceedings of the Second International Workshop on Multimedia Privacy and Security (pp. 81–87).

Taylor, A. (2016). Before "fake news," there was soviet "disinformation". *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/worldviews/wp/2016/11/26/before-fake-news-there-was-soviet-disinformation/

The Trust Project(2017). Retrieved from https://thetrustproject.org/.

Timberg, C., & Dwoskin, E. (2018, July). Twitter is sweeping out fake accounts like never before, Putting user growth at risk. *Washington Post*. Retrieved from https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/

Timsit, A. (2019, February 12). In the age of fake news, here's how schools are teaching kids to think like fact-checkers. *Quartz*. Retrieved from https://qz.com/1533747/in-the-age-of-fake-news-heres-how-schools-are-teaching-kids-to-think-like-fact-checkers/

Tong, A., Du, D.-Z., & Wu, W. (2018). On misinformation containment in online social networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31* (pp. 341–351). Red Hook, NY: Curran Associates Inc Retrieved from http://papers.nips.cc/paper/7317-on-misinformation-containment-in-online-social-networks.pdf

Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., & Krause, A. (2018). *Fake news detection in social networks via crowd signals*. In Companion Proceedings of the web Conference 2018, International World Wide Web Conferences Steering Committee. pp. 517–524.

Tugend, A. (2020). These students are learning about fake news and how to spot it. *The New York Times*. Retrieved from https://www.nytimes.com/2020/02/20/education/learning/news-literacy-2016-election.html

van der Tempel, J., & Alcock, J. E. (2015). Relationships between conspiracy mentality hyperactive agency detection, and schizotypy: Supernatural forces at work? *Personality and Individual Differences*, *82*, 136–141.

Van Duyn, E., & Collier, J. (2019). Priming and fake news: The effects of elite discourse on evaluations of news media. *Mass Communication and Society*, *22*(1), 29–48.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). *Online human-bot interactions*: *Detection, estimation, and characterization*. In ICWSM. pp. 280–289.

Vedova, M. L. D., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., & de Alfaro, L. (2018). *Automatic Online Fake News Detection Combining Content and Social Signals*. In 2018 22nd Conference of Open Innovations Association (FRUCT). IEEE. Retrieved from https://doi.org/10.23919%2Ffruct.2018.8468301.

Vo, N., & Lee, K. (2018). The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 275–284). ACM.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Vuković, M., Pripužić, K., & Belani, H. (2009). An intelligent automatic hoax detection system. In *Knowledge-based and intelligent information and engineering systems* (pp. 318–325). Berlin/Heidelberg: Springer Retrieved from https://doi.org/10.1007%2F978-3-642-04595-0_39

Waddington, K. (2012). *Gossip and organizations*. London: Routledge.

Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M., Zheng, H., & Zhao, B. Y. (2013). Social turing tests: Crowdsourcing sybil detection. *arXiv preprint arXiv:1205.3856*. Internet Society.

Wang, W. Y. (2017). Liar, liar pants on fire: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., ... Gao, J. (2018). *EANN: Event adversarial neural networks for multi-modal fake news detection*. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 849–857.

Wu, K., Yang, S., & Zhu, K. Q. (2015). *False rumors detection on sina weibo by propagation structures*. In 2015 IEEE 31st International Conference on Data Engineering. pp. 651–662.

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media. *ACM SIGKDD Explorations Newsletter*, *21*(2), 80–90 Retrieved from https://doi.org/10.1145%2F3373464.3373475

Xie, Y., Yu, F., Achan, K., Panigrahy, R., Hulten, G., & Osipkov, I. (2008). Spamming botnets: Signatures and characteristics. *ACM SIGCOMM Computer Communication Review*, *38*(4), 171–182.

Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., & Liu, H. (2019). Unsupervised fake news detection on social media: A generative approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 5644–5651 Retrieved from https://doi.org/10.1609%2Faaai.v33i01.33015644

Yang, X., Li, Y., & Lyu, S. (2019). *Exposing deep fakes using inconsistent head poses*. In ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8261–8265.

Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., & Yu, P. S. (2018). TI-CNN: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749*.

Yin, X., Han, J., & Philip, S. Y. (2008). Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, *20*(6), 796–808.

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Information diffusion in social media. In *Social media mining* (pp. 179–214). Cambridge, MA: Cambridge University Press Retrieved from https://doi.org/10.1017%2Fcbo9781139088510.008

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 9051–9062). Red Hook, NY: Curran Associates Inc.

Zhang, C. M., & Paxson, V. (2011). Detecting and analyzing automated activity on twitter. In N. Spring & G. Riley (Eds.), *Passive and active measurement (PAM 2011), LNCS 6579* (pp. 102–111). Berlin: Springer.

Zhang, D. Y., Wang, D., & Zhang, Y. (2017). *Constraint-aware dynamic truth discovery in big data social media sensing*. In 2017 IEEE International Conference on Big Data (Big Data). pp. 57–66.

Zhang, X., Karaman, S., & Chang, S.-F. (2019). Detecting and simulating artifacts in gan fake images. *arXiv preprint arXiv:1907.06515*.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). *Unpaired image-to-image translation using cycle-consistent adversarial networks*. In Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232.

Zubiaga, A. (2018). Learning class-specific word representations for early detection of hoaxes in social media. *arXiv preprint arXiv:1801.07311*.

Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS One*, *11*(3), e0150989.

**How to cite this article:** Shu K, Bhattacharjee A, Alatawi F, et al. Combating disinformation in a social media age. *WIREs Data Mining Knowl Discov*. 2020;10:e1385. https://doi.org/10.1002/widm.1385

## APPENDIX: DATASETS AND TOOLS AVAILABLE

With the gradually increasing volume of ongoing research in the domain of fake news or disinformation detection, there are a considerable number of datasets for use by researchers.

### Datasets

There have been a few small fake news datasets predating the era during which research in fake news detection gained momentum. The use of advanced computational methods and machine learning frameworks engenders the need for

large datasets that can be used to train such automatic detection models to achieve improved performance. Here we describe a few datasets that have been in use recently:

**[Twitter Media Corpus[3]]**: This dataset consists of tweets from Twitter containing multimedia content and associated "fake" or "real" labels. Tweets where such fake images were used as supporting media have been labeled as fake. For more information on the dataset and the framework developed by the creators of the dataset, please refer to Boididou et al. (2018).

**[PHEME Rumor Dataset[4]]**: This dataset (Kochkina, Liakata, & Zubiaga, 2018) contains a collection of tweets from Twitter, comprising of both rumors and non-rumors. These tweets correspond to nine different events, and are annotated with labels—True, False, or Unverified. For details of the data collection and the related analysis by the authors, please refer to Zubiaga, Liakata, Procter, Hoi, and Tolmie (2016).

**[CREDBANK[5]]**: This is a dataset (Mitra & Gilbert, 2015) containing over 60 million tweets from Twitter, over 1,049 real-world events. As explained in the paper introducing this dataset, each of the tweets have been annotated with a credibility score as judged by human annotator. The credibility score is based on a credibilty scale: *Certainly Accurate*, *Probably Accurate*, *Uncertain*, *Probably Inaccurate*, and *Certainly Inaccurate*.

**[LIAR[6]]**: Consists of 12.8K labeled sentences across different contexts, extracted from POLITIFACT.COM. The sentences consist of statements made by speakers having different political affiliations and also Facebook posts. Each statement is associated with a fine-grained truthfulness label, which is one of six possible values. Furthermore, each speaker is associated with a "credit history," that is a tuple having counts of how many statements they made in each of the truthfulness category, thus serving as some sort of credibility score. More information regarding the dataset can be found in Wang (2017).

**[Twitter Death Hoax[7]]**: This dataset consists of death reports posted on Twitter between January 1, 2012 and December 31, 2014. There are over 4k reports, out of which only 2031 are real deaths, as verified by using information from Wikipedia. For more details on the collection and related hoax detection task, please refer to Zubiaga (2018).

**[FA-KES: A Fake News Dataset around the Syrian War[8]]**: This dataset consists of news articles focused on the Syrian war and these articles are labeled as "fake" or "credible" by comparing the content of the articles with ground truth obtained from Syrian Violations Documentation Center. For more information regarding the data collection and annotation process, please refer to Salem, Al Feel, Elbassuoni, Jaber, and Farah (2019).

**[FakeNewsNet[9]]**: Consists of real and fake news articles related to politics (from POLITIFACT and celebrities (from GOSSIPCOP. Along with the content of the posts/tweets, the dataset also has data related to the social context - including user-profiles, user-following, user-followers, and so on. For more details, please refer to Shu, Mahudeswaran, Wang, Lee, and Liu (2018).

## Tools

Given the significant amount of ongoing work in the field of fake news or disinformation detection, researchers and practitioners have made certain tools available for use, to further the research in this area. Here, we list a few publicly available ones:

**[Big Bird[10]]**: This is a tool to automatically generate (fake) news articles. The website www.notrealnews.net contain "news" generated by this text generator tool and subsequently edited by human agents.

**[Computational-Verification[11]]**: This is a framework for identifying the veracity of content/posts on social media (Twitter, in this case). This makes sue of two types of features—*Tweet-based features*, such as number of words, number of retweets, and so on, and *User-based features* such as friend-follower ratio, number of tweets, and so on. A two-level classification model, to identify fake versus real content, is trained. For more details and information on the approach, please refer to Boididou et al. (2018).

**[Facebook Hoax[12]]**: This is a collection of Python scripts to collect posts on a page on Facebook, since a predefined date as input by the user. Authors in Tacchini, Ballarin, Della Vedova, Moret, and de Alfaro (2017) use this to create a dataset of over 15,000 posts on Facebook and try to identify hoax versus non-hoax posts based on the users who "liked" the posts. The authors used logistic regression and harmonic boolean label crowdsourcing to perform the classification. For more details, please refer to Tacchini et al. (2017).